Technical Report Machine Learning

"Breast Cancer Dataset Exploration Using Python and Sckit-Learn"

Telkom University

Disusun oleh :

Nurrafi Bagus Pratama

1103180026

Program Studi Teknik Komputer

Fakultas Teknik Elektro

Universitas Telkom

2023

Pendahuluan

Kanker payudara adalah kondisi di mana sel-sel di dalam payudara berkembang menjadi abnormal dan mulai tumbuh secara tidak terkontrol, membentuk tumor ganas atau kanker. Kanker payudara merupakan salah satu jenis kanker yang paling umum terjadi pada wanita, meskipun pada beberapa kasus juga dapat terjadi pada pria. Pada Technical report ini kita akan melakukan penelitian terhadap dataset kanker payudara menggunakan puython dan sckit-learn.

Deskripsi Dataset

Dataset ini berisi informasi tentang sampel tumor payudara dari pasien yang menderita kanker payudara serta informasi mengenal ukuran dan bentuk tumornya. Dataset ini terdiri dari 569 sampel dengan 30 fitur. Salah satu fiturnya yaitu ‘target’ yang merupakan variable biner yang menunjukkan keganasan tumor.

Visualisasi Data

Visualisasi data adalah proses membuat grafik atau representasi visual lainnya dari data numerik atau kualitatif untuk membantu memahami dan menganalisis pola, tren, atau hubungan di antara data tersebut. Tujuan utama dari visualisasi data adalah untuk membantu kita memahami dan menganalisis informasi dengan lebih mudah dan cepat, serta membuat keputusan yang lebih baik.Kali ini saya menggunakan Scatterplot dan histplot.

Scatterplot adalah jenis grafik yang menunjukkan hubungan antara dua variabel numerik. Grafik ini terdiri dari titik-titik yang ditempatkan di koordinat kartesian, dengan sumbu x mewakili satu variabel dan sumbu y mewakili variabel lainnya. Histplot atau histogram adalah jenis grafik yang menampilkan distribusi frekuensi dari suatu variabel numerik. Histogram mengelompokkan data menjadi beberapa kelas atau interval dan
menunjukkan jumlah pengamatan atau frekuensi pada setiap kelas.

Decision Tree

Decision tree atau pohon keputusan adalah model prediksi yang menggambarkan sebuah pohon keputusan dengan menggunakan algoritma pembelajaran mesin. Model ini digunakan untuk memecahkan masalah klasifikasi dan regresi, dengan mengambil keputusan berdasarkan serangkaian aturan dan kriteria yang didefinisikan pada setiap cabang atau simpul pada pohon keputusan.

Setiap simpul pada pohon keputusan mewakili sebuah kondisi atau kriteria yang digunakan untuk membagi data menjadi dua atau lebih kelompok, sedangkan setiap cabang pada simpul tersebut mewakili pilihan atau keputusan yang diambil berdasarkan kondisi atau kriteria tersebut. Pada akhirnya, prediksi dilakukan dengan melewati pohon keputusan hingga mencapai daun atau simpul akhir yang menunjukkan kelas atau nilai target.

Random Forrest

Random forest adalah model machine learning yang menggunakan sejumlah besar pohon keputusan atau decision tree dalam satu model. Random forest menggabungkan kekuatan dari banyak pohon keputusan, sehingga dapat menghasilkan prediksi yang lebih akurat dan stabil dibandingkan dengan satu pohon keputusan tunggal.

Setiap pohon keputusan dalam random forest dibangun secara acak, dengan memilih subset acak dari variabel input dan pengamatan acak dari dataset training. Hal ini membantu mengurangi overfitting dan meningkatkan variasi pada model, sehingga meningkatkan performa prediksi.

Self Training

Self-training adalah teknik pembelajaran mesin semi-supervised yang melibatkan dua tahap utama. Tahap pertama adalah pembuatan model pembelajaran mesin menggunakan dataset kecil yang telah dilabeli atau diklasifikasikan dengan benar. Tahap kedua melibatkan penggunaan model yang dibuat pada tahap pertama untuk membuat prediksi pada dataset yang belum dilabeli, dan kemudian memilih beberapa prediksi yang paling dapat diandalkan untuk ditambahkan ke dataset pelatihan.